

# Polytope conditioning and linear convergence of the Frank-Wolfe algorithm

Javier Peña\*      Daniel Rodríguez†

April 27, 2016

## Abstract

It is known that the gradient descent algorithm converges linearly when applied to a strongly convex function with Lipschitz gradient. In this case the algorithm's rate of convergence is determined by condition number of the function. In a similar vein, it has been shown that a variant of the Frank-Wolfe algorithm with away steps converges linearly when applied a strongly convex function over a polytope. In a nice extension of the unconstrained case, the algorithm's rate of convergence is determined by the product of the condition number of the function and a certain condition number of the polytope.

We shed new light into the latter type of polytope conditioning. In particular, we show that previous and seemingly different approaches to define a suitable condition measure for the polytope are essentially equivalent to each other. Perhaps more interesting, they can all be unified via a parameter of the polytope that formalizes a key premise linked to the algorithm's linear convergence. We also give new insight into the linear convergence property. For a convex quadratic objective, we show that the rate of convergence is determined by the condition number of a suitably scaled polytope.

---

\*Tepper School of Business, Carnegie Mellon University, USA, [jfp@andrew.cmu.edu](mailto:jfp@andrew.cmu.edu)

†Department of Mathematical Sciences, Carnegie Mellon University, USA, [drod@cmu.edu](mailto:drod@cmu.edu)

# 1 Introduction

It is a standard result in convex optimization that the gradient descent algorithm converges linearly to the minimizer of a strongly convex function with Lipschitz gradient. For a related discussion, e.g., [13, Chapter 2] or [4, Chapter 1]. Furthermore, in this case the rate of convergence is determined by the *condition number* of the objective function, that is, the ratio between the Lipschitz parameter of the gradient and the strong convexity parameter of the function.

In analogous fashion, the Frank-Wolfe algorithm [6, 9], also known as conditional gradient algorithm, for the problem  $\min_{y \in C} f(y)$  converges linearly to the minimizer of  $f$  on a compact convex set  $C$  provided  $f$  is strongly convex and the optimal solution lies in the relative interior of  $C$ . For a related discussion see, e.g., [3, 5, 9, 11] and the references therein. The assumption that the optimal solution belong to relative interior of  $C$  is critical for the linear convergence of the algorithm. Indeed, the rate of convergence depends on how far the optimal solution is from the relative boundary of  $C$ . In particular, this rate deteriorates when the optimal solution is near the relative boundary of  $C$ , and linear convergence breaks down altogether when the optimal solution is on the relative boundary of  $C$ .

When  $C = \text{conv}(A)$  for a finite set of atoms  $A$  and an oracle  $g \mapsto \text{argmin}_{a \in A} \langle g, a \rangle$  is available, Wolfe [15] proposed a variant of the Frank-Wolfe algorithm that includes *away steps*. Several articles have shown linear convergence results for this kind of variant of the Frank-Wolfe algorithm, including [1, 8, 7, 10, 12] as well as the more recent articles [2, 11, 14]. A common feature of [2, 11, 14] is that the statement of linear convergence is given in terms of the condition number of the objective function  $f$  and some type of *condition number* of the polytope  $\text{conv}(A)$ .

As we explain in Section 3, a generic version of linear convergence as in [2, 11, 14] hinges on three main premises. The first premise is a certain *slope bound* on the objective function and its optimal solution set. This first premise readily holds for strongly convex functions as it does in the unconstrained case. The second premise is a *decrease condition* on the objective function at each iteration of the algorithm. As in the unconstrained case, the second premise holds as long as an upper bound on the Lipschitz constant of the gradient is available, or if an appropriate line-search is performed at each iteration. The third premise, which seems to be the main technical component in each of the papers [2, 11, 14], is a premise on the search direction selected by the algorithm at each iteration. Loosely speaking, this third premise is a condition on the alignment of the algorithm's search direction with the negative of the gradient at the current iterate. The premise is that

this alignment should be comparable to that of a direct step towards the optimal solution. Unlike the first two premises, that essentially match the premises leading to the linear convergence of gradient descent in the unconstrained case, the third premise is inherent to the polytope defining the constraint set of the problem. This third premise can be formalized in terms of a certain kind of *condition number* of the polytope. In a nice extension of the unconstrained case, the rate of linear convergence of the Frank-Wolfe algorithm with away steps is determined by the product of the usual condition number of the objective function and the condition number of the polytope. (See Theorem 4 in Section 3.)

The central goal of this paper is to shed new light into this polytope’s condition number. The three articles [2, 11, 14] make different attempts to define a suitable polytope’s condition measure along the lines of the third premise sketched above. Each of these attempts has different merits and limitations. One of this paper’s main contributions is to show that these three kinds of condition measures, namely the *pyramidal width* defined by Lacoste-Julien and Jaggi [11], the *vertex-facet distance* defined by Beck and Shtern [2], and the *restricted width* defined by Peña, Rodríguez, and Soheili [14], turn out to be essentially equivalent. Perhaps more important, they are all unified via a *facial distance* of the polytope. As we explain in Section 2 and Section 3, the facial distance can be seen as a natural quantity associated to the polytope that formalizes a key alignment condition of the search direction at each iteration of the Frank-Wolfe algorithm with away steps.

Section 2 presents the technical bulk of our paper. One of our results (Theorem 1) is a characterization of the facial distance of a polytope as the minimum distance between a face of the polytope and a kind of *complement polytope*. This characterization can be seen as a refinement of the *vertex-facet distance* proposed by Beck and Shtern [2]. Theorem 1 motivates the name “facial distance” for this quantity. Theorem 1 provides a method to compute or bound the facial distance as we illustrate in a few examples. We also show (Theorem 2) that the facial distance coincides with the pyramidal width defined by Lacoste-Julien and Jaggi [11]. As a byproduct of this result, we obtain a simplification of the original definition of pyramidal width. We also give a *localized* version of Theorem 1 for a kind of *localized* version of the facial distance of the polytope (Theorem 3).

As mentioned above, Section 3 details how the linear convergence of the Frank-Wolfe algorithm with away steps can be derived from three central premises. The goal of Section 3 is to highlight the role of these three key premises, particularly the third one. We discuss how the third premise is naturally tied to the facial distance of the polytope discussed in Section 2. Our exposition allows us to distill a key tradeoff in the existing bounds on the rate of convergence of the Frank-Wolfe algorithm

with away steps. On the one hand, the algorithm's rate of convergence can be bounded in terms of quantities that depend *only* on properties of the polytope and of the objective function but not on the optimal solution. More precisely, for a strongly convex objective function with Lipschitz gradient the rate of convergence can be bounded in terms of the product of the polytope's condition number and the objective function's condition number. This is a feature of the results in [11] but not of those in [2, 14] that depend on the optimal solution set. On the other hand, a *sharper* bound on the rate of convergence can be given if we allow it to depend on the location of the optimal solution in the polytope. More precisely, the rate of convergence can be bounded in terms of the product of a *localized* polytope's condition number that depends on the solution set and the objective function's condition number. The statement of Theorem 4 makes the connection between the two bounds completely transparent: The solution-independent bound is simply the most conservative solution-dependent one. Not surprisingly, the solution-dependent bound can be arbitrarily better than the solution-independent one.

Section 4 discusses the linear convergence property in the special but important case when the objective function is of the form  $f(y) = \frac{1}{2} \langle y, Qy \rangle + \langle b, y \rangle$  for  $Q$  positive semidefinite. As Theorem 5 in Section 4 shows, in this case the rate of convergence is determined by the condition number of the scaled polytope  $\text{conv}(Q^{1/2}A)$ . In Section 5 we show that the latter result extends, under suitable assumptions on the algorithm's choice of steplength, to a composite objective function of the form  $f(y) = g(Ey) + \langle b, y \rangle$  where  $g$  is a strongly convex function with Lipschitz gradient and  $E$  is a matrix of suitable size. (See Theorem 6 in Section 5.) This result is along the same lines of the linear convergence result of Beck and Shtern's [2, Theorem 3.1]. However, our bound on the rate of convergence and proof technique are fairly different.

## 2 The facial distance $\Phi(A)$

This section constitutes the technical bulk of the paper. We define the *facial distance*  $\Phi(A)$  and prove several interesting results about it. In particular, we show that it essentially matches the various kinds of condition measures previously defined in [2, 11, 14].

Assume  $A = \{a_1, \dots, a_n\} \subseteq \mathbb{R}^m$  is a finite set of atoms in  $\mathbb{R}^m$ . For convenience we will make the following slight abuse of notation: We will write  $A$  to denote both the set  $\{a_1, \dots, a_n\}$  and the matrix  $\begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ . The appropriate meaning of  $A$  will be clear from the context. Let  $\Delta_{n-1} := \{x \in \mathbb{R}_+^n : \|x\|_1 = 1\}$ . For  $x \in \Delta_{n-1}$ ,

define  $I(x) \subseteq \{1, \dots, n\}$  and  $S(x) \subseteq A$  as

$$I(x) := \{i \in \{1, \dots, n\} : x_i > 0\}$$

and

$$S(x) := \{a_i : i \in I(x)\}.$$

Observe that the sets  $I(x), S(x)$  define the *support* of  $Ax$ .

Throughout the paper,  $\|\cdot\|$  will denote the Euclidean norm. Assume  $x, z \in \Delta_{n-1}$  are such that  $A(x-z) \neq 0$  and let  $d := \frac{A(x-z)}{\|A(x-z)\|}$ . Define

$$\Phi(A, x, z) = \min_{p \in \mathbb{R}^m : \langle p, d \rangle = 1} \left\{ \max_{s \in S(x), a \in A} \langle p, s - a \rangle \right\}, \quad (1)$$

and

$$\Phi(A) = \min_{x, z \in \Delta_{n-1} : A(x-z) \neq 0} \Phi(A, x, z).$$

The connection between  $\Phi(A, x, z)$  and the Frank-Wolfe with away steps algorithm will be made explicit as Premise 3 in Section 3 but the basic idea is as follows. At each iteration the algorithm starts from a current point  $y = Ax \in \text{conv}(A)$  and selects the two atoms  $a, s \in A$  that attain  $\max_{s \in S(x), a \in A} \langle p, s - a \rangle$  for  $p = -\nabla f(y)$ . Premise 3 requires that for  $d := \frac{A(x-z)}{\|A(x-z)\|}$  the ratio  $\frac{\langle p, s-a \rangle}{\langle p, d \rangle}$  be bounded away from zero. The latter condition means the alignment of the vector  $a - s$  and the direction  $p$  should be comparable to the alignment of  $d$  and  $p$ . The need to formalize this premise motivates the definition of the quantities  $\Phi(A, x, z)$  and  $\Phi(A)$ .

Notice the asymmetry between the roles  $x$  and  $z$  in  $\Phi(A, x, z)$ . We can think of  $z$  as defining an *anchor* point  $Az \in \text{conv}(A)$ . When  $Az = 0$ , the quantity  $\Phi(A, x, z)$  coincides with the quantity  $\phi(A, x)$  defined in [14]. Thus  $\Phi(A)$  can be seen as a refinement of the *restricted width*  $\phi(A) = \min_{x \in \Delta_{n-1} : Ax \neq 0} \phi(A, x)$  defined in [14]. The quantity  $\phi(A, x)$  was introduced in a form more closely related to the alternative expression (2) for  $\Phi(A, x, z)$  in Proposition 1 below. The expression (2) characterizes  $\Phi(A, x, z)$  as the length of the longest segment in  $\text{conv}(A)$  in the direction  $A(x-z)$  with one endpoint in  $\text{conv}(S(x))$  and the other in  $\text{conv}(A)$ .

**Proposition 1** Assume  $A = \{a_1, \dots, a_n\} \subseteq \mathbb{R}^m$ ,  $x, z \in \Delta_{n-1}$  are such that  $A(x-z) \neq 0$  and let  $d := \frac{A(x-z)}{\|A(x-z)\|}$ . Then

$$\Phi(A, x, z) = \max_{w, y, \lambda} \{ \lambda > 0 : \exists w, y \in \Delta_{n-1}, I(w) \subseteq I(x), A(w-y) = \lambda d \}. \quad (2)$$

Furthermore, if  $p$  attains the minimum value  $\Phi(A, x, z)$  in (1) and  $u = Aw, v = Ay$  maximize the right hand side in (2) then  $v \in \text{conv}(B)$  and  $u \in \text{conv}(A \setminus B)$  where  $B = \text{Argmin}_{a \in A} \langle p, a \rangle$ .

**Proof:** To ease notation, let  $I := I(x)$ . Observe that the right-hand-side in (2) is

$$\begin{aligned} \max_{w_I, y, \lambda} \quad & \lambda \\ & A_I w_I - A y - \lambda d = 0 \\ & \mathbf{1}_I^T w_I = 1 \\ & \mathbf{1}^T y = 1 \\ & w_I, y \geq 0. \end{aligned}$$

On the other hand, from the definition (1) of  $\Phi(A, x, z)$ , it follows that

$$\begin{aligned} \Phi(A, x, z) = \min_{p, t, \tau} \quad & t + \tau \\ & A_I^T p \leq t \mathbf{1}_I \\ & A^T p \geq -\tau \mathbf{1} \\ & \langle d, p \rangle = 1. \end{aligned}$$

Therefore (2) and the subsequent statement follow by linear programming duality.  $\blacksquare$

Theorem 1 below gives a characterization of  $\Phi(A)$  in terms of the minimum distance between a face  $F$  of  $\text{conv}(A)$  and its *complement polytope*  $\text{conv}(A \setminus F)$ . This characterization motivates the name *facial distance* for the quantity  $\Phi(A)$ . The minimum distance expression for  $\Phi(A)$  in Theorem 1 can be seen as a refinement of the so-called *vertex-facet distance constant* defined by Beck and Shtern [2].

**Theorem 1** *Assume  $A = \{a_1, \dots, a_n\} \subseteq \mathbb{R}^m$  and at least two of these points are different. Then*

$$\Phi(A) = \min_{\substack{F \in \text{faces}(\text{conv}(A)) \\ \emptyset \subsetneq F \subsetneq \text{conv}(A)}} \text{dist}(F, \text{conv}(A \setminus F)).$$

*Furthermore, if  $F \in \text{faces}(\text{conv}(A))$  minimizes the right hand side, then there exist  $x, z \in \Delta_{n-1}$  such that  $Az \in F, Ax \in \text{conv}(A \setminus F)$  and*

$$\Phi(A) = \Phi(A, x, z) = \max_{s \in S(x), a \in A} \langle p, s - a \rangle = \|A(x - z)\|$$

$$\text{for } p = \frac{A(x - z)}{\|A(x - z)\|}.$$

**Proof:** We first show that  $\Phi(A) \geq \min_{F \in \text{faces}(\text{conv}(A))} \text{dist}(F, \text{conv}(A \setminus F))$ .

To that end, assume  $x, z \in \Delta_{n-1}$  are such that  $A(x - z) \neq 0$  and let  $d := \frac{A(x - z)}{\|A(x - z)\|}$ . Let  $p \in \mathbb{R}^m$  be a vector attaining the minimum in (1). Consider the face  $F$  of  $\text{conv}(A)$  defined as

$$F = \underset{v \in \text{conv}(A)}{\text{Argmin}} \langle p, v \rangle.$$

From Proposition 1 it follows that  $\Phi(A, x, z) = \|u - v\|$  for some  $v \in F$  and  $u \in \text{conv}(A \setminus F)$ . Therefore

$$\text{dist}(F, \text{conv}(A \setminus F)) \leq \|u - v\| = \Phi(A, x, z).$$

Since this holds for any  $x, z \in \Delta_{n-1}$  such that  $A(x - z) \neq 0$ , it follows that  $\Phi(A) \geq \min_{F \in \text{faces}(\text{conv}(A))} \text{dist}(F, \text{conv}(A \setminus F))$ .

Next we show that  $\Phi(A) \leq \min_{F \in \text{faces}(\text{conv}(A))} \text{dist}(F, \text{conv}(A \setminus F))$ . To that end, assume  $F$  is a face of  $\text{conv}(A)$  that minimizes  $\text{dist}(F, \text{conv}(A \setminus F))$ . Let  $u \in \text{conv}(A \setminus F)$  and  $v \in F$  be such that

$$\text{dist}(F, \text{conv}(A \setminus F)) = \|u - v\|. \quad (3)$$

By taking a face of  $F$  if necessary, we can assume that  $v$  is in the relative interior of  $F$ . Likewise, we can assume that  $u$  is in the relative interior of some face  $G$  of  $\text{conv}(A \setminus F)$ . From (3) it follows that  $u - v$  is orthogonal to  $F$  and  $G$ , that is,

$$\langle u - v, s - u \rangle = \langle u - v, t - v \rangle = 0 \quad \text{for all } t \in F, s \in G. \quad (4)$$

Next we claim that

$$\langle u - v, a - v \rangle \geq 0 \quad \text{for all } a \in A. \quad (5)$$

We prove this claim by contradiction. Assume  $a \in A$  is such that  $\langle u - v, a - v \rangle < 0$ , then  $\langle u - v, a - u \rangle = \langle u - v, a - v \rangle - \|u - v\|^2 < 0$  and from (4) we get  $a \notin F$ . Hence for  $\lambda > 0$  sufficiently small the point  $u + \lambda(a - u) \in \text{conv}(A \setminus F)$  satisfies

$$\|u + \lambda(a - u) - v\|^2 = \|u - v\|^2 + 2\lambda \langle u - v, a - u \rangle + \lambda^2 \|v\|^2 < \|u - v\|^2,$$

which contradicts (3). Thus (5) is proven.

Let  $x, z \in \Delta_{n-1}$  be such that  $u = Ax, v = Az$  and  $S(x) = A \cap G$ . The latter is possible since  $u \in \text{conv}(G)$ . We finish by observing that for  $p = d = \frac{A(x-z)}{\|A(x-z)\|} = \frac{u-v}{\|u-v\|}$

$$\begin{aligned} \Phi(A, x, z) &\leq \max_{s \in S(x), a \in A} \langle p, s - a \rangle \\ &= \max_{s \in G, t \in \text{conv}(A)} \langle p, s - t \rangle \\ &= \langle p, u - v \rangle \\ &= \|u - v\| \\ &= \text{dist}(F, \text{conv}(A \setminus F)). \end{aligned}$$

The third step follows from (4) and (5). The fourth step follows from (3). Next, observe that  $\text{dist}(F, \text{conv}(A \setminus F)) \leq \Phi(A) \leq \Phi(A, x, z)$

and hence we have equality in the first step above and all of the expressions at each step are equal to  $\Phi(A)$ .  $\blacksquare$

From Theorem 1 we can readily compute the values of  $\Phi(A)$  in the special cases detailed in the examples below. We use the following notation. Let  $e \in \mathbb{R}^m$  denote the vector with all components equal to one, and for  $i = 1, \dots, m$  let  $e_i \in \mathbb{R}^m$  denote the vector with  $i$ -th component equal to one and all others equal to zero.

**Example 1** Suppose  $A = \{0, 1\}^m \subseteq \mathbb{R}^m$ . By Theorem 1, induction, and symmetry it follows that

$$\Phi(A) = \text{dist}(0, \text{conv}(A \setminus \{0\})) = \text{dist}(0, \text{conv}\{e_1, \dots, e_m\}) = \frac{\|e\|}{m} = \frac{1}{\sqrt{m}}.$$

**Example 2** Let  $A = \{e_1, \dots, e_m\} \subseteq \mathbb{R}^m$ . By Theorem 1 and the facial structure of  $\text{conv}(A)$  it follows that

$$\begin{aligned} \Phi(A) &= \min_{\emptyset \subsetneq S \subsetneq A} \text{dist}(\text{conv}(S), \text{conv}(A \setminus S)) \\ &= \min_{\emptyset \subsetneq S \subsetneq A} \left\| \frac{\sum_{s \in S} s}{|S|} - \frac{\sum_{a \in A \setminus S} a}{|A \setminus S|} \right\| \\ &= \min_{r \in \{1, \dots, m-1\}} \sqrt{\frac{m}{r(m-r)}} \\ &= \begin{cases} \frac{2}{\sqrt{m}} & \text{if } m \text{ is even} \\ \frac{2}{\sqrt{m-\frac{1}{m}}} & \text{if } m \text{ is odd.} \end{cases} \end{aligned}$$

We note that the values for  $\Phi(A)$  in the above examples match exactly the values of the *pyramidal width* defined by Lacoste-Julien and Jaggi [11]. Theorem 2 below shows that indeed the pyramidal width is identical to the facial distance  $\Phi(A)$ . As a byproduct of this identity, the original definition of pyramidal width given in [11] can be simplified.

Lacoste-Julien and Jaggi define the *pyramidal directional width* of a finite set  $A \subseteq \mathbb{R}^m$  with respect to a direction  $r \in \mathbb{R}^m$  and a base point  $u \in \text{conv}(A)$  as

$$\text{PdirW}(A, r, u) := \min_{S \in S_u} \max_{a \in A, s \in S} \left\langle \frac{r}{\|r\|}, a - s \right\rangle$$

where  $S_u = \{S \subseteq A : u \in \text{conv}(S)\}$ . Lacoste-Julien and Jaggi also define the *pyramidal width* of a set  $A$  as

$$\text{PWidth}(A) := \min_{\substack{F \in \text{faces}(\text{conv}(A)) \\ u \in F, r \in \text{cone}(F - u) \setminus \{0\}}} \text{PdirW}(F \cap A, r, u).$$



Observe that  $r \in \text{cone}(F-u) \setminus \{0\}$  if and only if  $r$  is a positive multiple of some  $v-u$  where  $u, v \in F$  and  $u-v \neq 0$ . Therefore

$$\text{PWidth}(A) = \min_{\substack{F \in \text{faces}(\text{conv}(A)) \\ u, v \in F, u \neq v}} \text{PdirW}(F \cap A, v-u, u). \quad (6)$$

**Theorem 2** Assume  $A = \{a_1, \dots, a_n\} \subseteq \mathbb{R}^m$  and at least two of these points are different. Then

$$\Phi(A) = \min_{u, v \in \text{conv}(A), u \neq v} \text{PdirW}(A, v-u, u) = \text{PWidth}(A). \quad (7)$$

**Proof:** Assume  $F \in \text{faces}(A)$  minimizes  $\text{dist}(F, \text{conv}(A \setminus F))$  and  $x, z \in \Delta_{n-1}$  are as in Theorem 1. Then for  $p = \frac{A(x-z)}{\|A(x-z)\|}$

$$\Phi(A) = \max_{s \in S(x), a \in A} \langle p, s-a \rangle.$$

Therefore for  $u = Ax$  and  $v = Az$  we have  $S(x) \in S_u$  and

$$\begin{aligned} \Phi(A) &= \max_{a \in A, s \in S(x)} \left\langle \frac{v-u}{\|v-u\|}, a-s \right\rangle \\ &\geq \min_{S \in S_u} \max_{a \in A, s \in S} \left\langle \frac{v-u}{\|v-u\|}, a-s \right\rangle \\ &= \text{PdirW}(A, v-u, u). \end{aligned}$$

Hence we conclude that  $\Phi(A) \geq \min_{u, v \in \text{conv}(A), u \neq v} \text{PdirW}(A, v-u, u)$ .

On the other hand, for  $u, v \in \text{conv}(A), u \neq v$  let  $S \in S_u$  be such that

$$\text{PdirW}(A, v-u, u) := \max_{a \in A, s \in S} \left\langle \frac{v-u}{\|v-u\|}, a-s \right\rangle.$$

Let  $x, z \in \Delta_{n-1}$  be such that  $Ax = u, Az = v$ . Since  $S \in S_u$ , we can assume that  $S(x) \subseteq S$ . Taking  $p = d = \frac{A(x-z)}{\|A(x-z)\|} = -\frac{v-u}{\|v-u\|}$  it follows that

$$\begin{aligned} \Phi(A) &\leq \Phi(A, x, z) \leq \max_{s \in S(x), a \in A} \langle p, s-a \rangle \\ &= \max_{a \in A, s \in S(x)} \left\langle \frac{v-u}{\|v-u\|}, a-s \right\rangle \\ &\leq \max_{a \in A, s \in S} \left\langle \frac{v-u}{\|v-u\|}, a-s \right\rangle \\ &= \text{PdirW}(A, v-u, u). \end{aligned}$$

Consequently  $\Phi(A) \leq \min_{u, v \in \text{conv}(A), u \neq v} \text{PdirW}(A, v-u, u)$  as well. Therefore the identity between the first two terms in (7) follows. To finish,

observe that by (6)

$$\begin{aligned}
\text{PWidth}(A) &= \min_{F \in \text{faces}(\text{conv}(A))} \min_{u, v \in \text{conv}(F), u \neq v} \text{PdirW}(F \cap A, v - u, u) \\
&= \min_{F \in \text{faces}(\text{conv}(A))} \Phi(F \cap A) \\
&= \Phi(A).
\end{aligned}$$

The second step follows by applying the identity between the first two terms in (7) to each  $F \cap A$ . The third step follows by Theorem 1.  $\blacksquare$

As we will see in Section 3 below, the following *localized* variant of  $\Phi(A)$ , which could be quite a bit larger than  $\Phi(A)$  plays a role in the linear convergence rate of the Frank-Wolfe algorithm. Assume  $A = \{a_1, \dots, a_n\} \subseteq \mathbb{R}^m$  and at least two of these points are different. For  $z \in \Delta_{n-1}$  let

$$\Phi(A, z) := \min_{x \in \Delta_{n-1} : A(x-z) \neq 0} \Phi(A, x, z),$$

and for  $Z \subseteq \Delta_{n-1}$  let

$$\Phi(A, Z) := \inf_{z \in Z} \Phi(A, z).$$

Theorem 3 gives a localized version of Theorem 1 for  $\Phi(A, Z)$ . Its proof relies on the following localized version of Proposition 1. We omit the proof of Proposition 2 as it is a straightforward modification of the proof of Proposition 1.

**Proposition 2** Assume  $A = \{a_1, \dots, a_n\} \subseteq \mathbb{R}^m$ ,  $x, z \in \Delta_{n-1}$  are such that  $A(x-z) \neq 0$  and let  $d := \frac{A(x-z)}{\|A(x-z)\|}$ . If  $F \in \text{faces}(\text{conv}(A))$  contains  $Az$  then

$$\begin{aligned}
\min_{p: \langle p, d \rangle = 1} \max_{s \in S(x), a \in F \cap A} \langle p, s - a \rangle = \\
\max \{ \lambda > 0 : \exists w, y \in \Delta_{n-1}, I(w) \subseteq I(x), Ay \in F, A(w-y) = \lambda d \}.
\end{aligned}$$

Furthermore, if  $p$  minimizes the left hand side and  $u = Aw, v = Ay$  maximize the right hand side then  $v \in \text{conv}(B)$  and  $u \in \text{conv}(A \setminus B)$  where  $B = \text{Argmin}_{a \in F \cap A} \langle p, a \rangle$ .

**Theorem 3** Assume  $A = \{a_1, \dots, a_n\} \subseteq \mathbb{R}^m$  and at least two of these points are different. Then for  $Z \subseteq \Delta_{n-1}$

$$\Phi(A, Z) \geq \min_{G \in \text{faces}(F)} \text{dist}(G, \text{conv}(A \setminus G))$$

where  $F$  is the smallest face of  $\text{conv}(A)$  containing  $AZ = \{Az : z \in Z\}$ .

**Proof:** This is a straightforward modification of the proof of Theorem 1. Observe that for  $x \in \Delta_{n-1}$ ,  $z \in Z$  and  $d = \frac{A(x-z)}{\|A(x-z)\|}$

$$\Phi(A, x, z) \geq \min_{p: \langle p, d \rangle = 1} \max_{s \in S(x), a \in F \cap A} \langle p, s - a \rangle.$$

Let  $p \in \mathbb{R}^m$  be a vector attaining the minimum in the above right hand side and consider the face  $G$  of  $F$  defined as

$$G = \operatorname{Argmin}_{v \in F} \langle p, v \rangle.$$

From Proposition 2 it follows that for some  $u \in \operatorname{conv}(A \setminus G)$  and  $v \in G$

$$\begin{aligned} \Phi(A, x, z) &\geq \min_{p: \langle p, d \rangle = 1} \max_{s \in S(x), a \in F \cap A} \langle p, s - a \rangle \\ &= \|u - v\| \\ &\geq \operatorname{dist}(G, \operatorname{conv}(A \setminus G)). \end{aligned}$$

Since this holds for all  $x \in \Delta_{n-1}$ ,  $z \in Z$ , we conclude that

$$\Phi(A, Z) \geq \min_{G \in \operatorname{faces}(F)} \operatorname{dist}(G, \operatorname{conv}(A \setminus G)).$$

■

It is evident from Theorem 1 and Theorem 3 that  $\Phi(A, Z)$  can be arbitrarily larger than  $\Phi(A)$ . In particular, this would be the case for  $Z = \{e_1\}$  if  $A$  consists of the isolated atom  $\{a_1\}$  and some other atoms  $\{a_2, \dots, a_n\}$  clustered together.

### 3 Frank Wolfe algorithm with away steps

We next present a fairly generic linear convergence result for a version of the Frank Wolfe algorithm with away steps. We emphasize that the statement in Theorem 4 can be found in or inferred from results already shown in [2, 11] albeit via more involved arguments. The goal of this section is to highlight three central premises that yield the proof of linear convergence, namely a first premise in the form of a *slope bound*, a second premise in the form of a *decrease condition*, and a third premise on the *search direction* selected by the algorithm. As we detail below, the third premise is naturally tied to the condition measures  $\Phi(A)$  and  $\Phi(A, Z)$  discussed in Section 2.

Assume  $A = \{a_1, \dots, a_n\} \subseteq \mathbb{R}^m$  and consider the problem

$$\min_{y \in \operatorname{conv}(A)} f(y) \tag{8}$$

where  $f : \operatorname{conv}(A) \rightarrow \mathbb{R}$  is a differentiable convex function.

We will rely on the following notation related to problem (8) throughout the rest of the paper. Let  $f^*, Y^*, Z^*$  be defined as

$$f^* := \min_{y \in \text{conv}(A)} f(y), \quad Y^* := \underset{y \in \text{conv}(A)}{\text{Argmin}} f(y), \quad Z^* := \{z \in \Delta_{n-1} : Az \in Y^*\}.$$

---

**Algorithm 1** Frank-Wolfe Algorithm with Away Steps

---

```

1: Pick  $x_0 \in \Delta_{n-1}$ ; put  $y_0 := Ax_0$ ;  $k := 0$ 
2: for  $k = 0, 1, 2, \dots$  do
3:    $j := \underset{i=1, \dots, n}{\text{argmin}} \langle a_i, \nabla f(y_k) \rangle$ ;  $\ell := \underset{i \in I(x_k)}{\text{argmax}} \langle a_i, \nabla f(y_k) \rangle$ 
4:   if  $\langle a_j - y_k, \nabla f(y_k) \rangle < \langle y_k - a_\ell, \nabla f(y_k) \rangle$  then (regular step)
5:      $v := a_j - y_k$ ;  $u := e_j - x_k$ ;  $\theta_{\max} := 1$ 
6:   else (away step)
7:      $v := y_k - a_\ell$ ;  $u := x_k - e_\ell$ ;  $\theta_{\max} := \frac{(x_k)_\ell}{1 - (x_k)_\ell}$ 
8:   end if
9:   choose  $\theta_k \in [0, \theta_{\max}]$ 
10:   $x_{k+1} := x_k + \theta_k u$ ;  $y_{k+1} := y_k + \theta_k v = Ax_{k+1}$ 
11: end for

```

---

Theorem 4 gives a fairly generic linear convergence result for Algorithm 1. This result hinges on the following three premises.

**Premise 1:** *There exists  $\mu > 0$  such that the objective function  $f$  satisfies the following bound: For all  $y \in \text{conv}(A)$  and  $y^* \in Y^*$  with  $y^* \neq y$*

$$\left\langle \nabla f(y), \frac{y - y^*}{\|y - y^*\|} \right\rangle \geq \sqrt{2\mu(f(y) - f^*)}.$$

Premise 1 readily holds if  $f$  is strongly convex with parameter  $\mu$ . In this case the optimality of  $y^*$ , the strong convexity of  $f$ , and the arithmetic-geometric inequality imply

$$\langle \nabla f(y), y - y^* \rangle \geq f(y) - f^* + \frac{\mu}{2} \|y - y^*\|^2 \geq \|y - y^*\| \sqrt{2\mu(f(y) - f^*)}.$$

From the error bound of Beck and Shtern [2, Lemma 2.1], it follows that Premise 1 also holds in the more general case when  $f(y) = g(Ey) + \langle b, y \rangle$  for some strongly convex function  $g$ .

**Premise 2:** *There exists  $L > 0$  such that the steplength  $\theta_k$  at each iteration of Algorithm 1 satisfies either  $\theta_k = \theta_{\max}$  and  $f(y_k + \theta_{\max} v) \leq f(y_k)$ , or*

$$f(y_k + \theta_k v) \leq f(y_k) - \frac{\langle \nabla f(y_k), v \rangle^2}{2L\|v\|^2}.$$

Premise 2 holds if  $\nabla f$  is Lipschitz with constant  $L$  and the steplength is computed via

$$\theta_k = \operatorname{argmin}_{\theta \in [0, \theta_{\max}]} \left\{ \theta \langle \nabla f(y_k), v \rangle + \frac{L \|v\|^2 \theta^2}{2} \right\} = \max \left\{ -\frac{\langle \nabla f(y_k), v \rangle}{L \|v\|^2}, \theta_{\max} \right\}.$$

This follows because the Lipschitz continuity of  $\nabla f$  implies that for all  $\theta \in \mathbb{R}$  with  $y_k + \theta v \in \operatorname{conv}(A)$

$$f(y_k + \theta v) \leq f(y_k) + \theta \langle \nabla f(y_k), v \rangle + \frac{L \theta^2}{2} \|v\|^2.$$

**Premise 3:** *There exists  $c > 0$  such that the search direction  $v$  selected in Step 5 or Step 7 of Algorithm 1 satisfies*

$$-\frac{\langle \nabla f(y_k), v \rangle}{\langle \nabla f(y_k), d \rangle} \geq c$$

for all  $d = \frac{y_k - y^*}{\|y_k - y^*\|}$  such that  $y^* \in Y^*$  and  $\langle \nabla f(y_k), d \rangle > 0$ .

Premise 3 holds at each iteration of the Frank-Wolfe Algorithm with Away Steps for  $c = \frac{\Phi(A)}{2}$  as well as for the sharper bound  $c = \frac{\Phi(A, Z^*)}{2}$ . To see this, observe that the choice between regular or away steps ensures

$$-\langle \nabla f(y_k), v \rangle \geq \frac{1}{2} \langle \nabla f(y_k), a_\ell - a_j \rangle > 0.$$

Let  $z \in Z^*$ . Taking  $d = \frac{A(x_k - z)}{\|A(x_k - z)\|} = \frac{y_k - y^*}{\|y_k - y^*\|}$ ,  $p = \frac{\nabla f(y_k)}{\langle \nabla f(y_k), d \rangle}$ , and using the construction of  $\Phi(A, x, z)$  we get

$$\frac{\langle \nabla f(y_k), a_\ell - a_j \rangle}{\langle \nabla f(y_k), d \rangle} = \max_{\ell \in I(x_k), a \in A} \langle p, a_\ell - a \rangle \geq \Phi(A, x_k, z) \geq \Phi(A, Z^*).$$

Hence

$$-\frac{\langle \nabla f(y_k), v \rangle}{\langle \nabla f(y_k), d \rangle} \geq \frac{\langle \nabla f(y_k), a_\ell - a_j \rangle}{2 \langle \nabla f(y_k), d \rangle} \geq \frac{\Phi(A, Z^*)}{2} \geq \frac{\Phi(A)}{2}.$$

**Theorem 4** *Assume  $x_0 \in \Delta_{n-1}$  in Step 1 of Algorithm 1 is a vertex of  $\Delta_{n-1}$ . If Premise 1, Premise 2, and Premise 3 hold then the sequence of points  $\{y_k : k = 0, 1, \dots\}$  generated by Algorithm 1 satisfies*

$$f(y_k) - f^* \leq (1 - r)^{k/2} (f(y_0) - f^*) \quad (9)$$

for  $r = \frac{\mu c^2}{L \cdot \operatorname{diam}(A)^2}$ . In particular, (9) holds for the solution-independent rate

$$r = \frac{\mu}{L} \cdot \frac{\Phi(A)^2}{4 \operatorname{diam}(A)^2}$$

as well as for the sharper, though solution-dependent, rate

$$r = \frac{\mu}{L} \cdot \frac{\Phi(A, Z^*)^2}{4\text{diam}(A)^2}.$$

**Proof:** This proof is an adaptation of the proofs in [2, 11, 14]. If  $\theta_k < \theta_{\max}$ , then Premise 2 yields

$$f(y_{k+1}) \leq f(y_k) - \frac{\langle \nabla f(y_k), v \rangle^2}{2L\|v\|^2}. \quad (10)$$

Premise 3 and Premise 1 in turn yield

$$\langle \nabla f(y_k), v \rangle^2 \geq \frac{c^2 \langle \nabla f(y_k), y_k - y^* \rangle^2}{\|y_k - y^*\|^2} \geq 2\mu c^2 (f(y_k) - f^*).$$

Plugging the last inequality into (10) we get

$$\begin{aligned} f(y_{k+1}) - f^* &\leq \left(1 - \frac{\mu c^2}{L\|v\|^2}\right) (f(y_k) - f^*) \\ &\leq \left(1 - \frac{\mu c^2}{L \cdot \text{diam}(A)^2}\right) (f(y_k) - f^*). \end{aligned}$$

This happens whenever  $\theta_k < \theta_{\max}$ . When  $\theta_k = \theta_{\max}$  we have  $f(y_{k+1}) - f^* \leq f(y_k) - f^*$  by Premise 2. To finish, it suffices to show that after  $N$  iterations, for at least  $N/2$  of them we have  $\theta_k < \theta_{\max}$ . To that end, we apply the following clever argument from [11]. Each time  $\theta_k = \theta_{\max}$  we have  $|I(x_{k+1})| \leq |I(x_k)| - 1$  and each time  $\theta_k < \theta_{\max}$  we have  $|I(x_{k+1})| \leq |I(x_k)| + 1$ . Since  $|I(x_0)| = 1$  and  $|I(x_k)| \geq 1$  for all  $x_k \in \Delta_{n-1}$ , it follows that after  $N$  iterations there must have been at least as many iterations with  $\theta_k < \theta_{\max}$  as there were with  $\theta_k = \theta_{\max}$ . ■

As we noted at the end of Section 2, the quantity  $\Phi(A, Z^*)$  in Theorem 4 can be arbitrarily better (larger) than  $\Phi(A)$ . Observe that the solution-independent bound is simply the most conservative solution-dependent one for all possible solution sets  $Z^* \subseteq \Delta_{n-1}$ .

## 4 Convex quadratic objective

We next specialize the linear convergence result in Theorem 4 to the case when the objective function is of the form

$$f(y) = \frac{1}{2} \langle y, Qy \rangle + \langle b, y \rangle,$$

for an  $m \times m$  symmetric positive semidefinite matrix  $Q$  and  $b \in \mathbb{R}^m$ . Consider problem (8) and Algorithm 1 for this objective function. In this case  $\theta_k$  in Step 9 of Algorithm 1 can be easily computed via the following exact line-search:

$$\begin{aligned} \theta_k &:= \operatorname{argmin}_{\theta \in [0, \theta_{\max}]} f(y_k + \theta v) \\ &= \begin{cases} \min \left\{ \theta_{\max}, -\frac{\langle v, Qy_k + b \rangle}{\langle v, Qv \rangle} \right\} & \text{if } \langle v, Qv \rangle > 0 \\ \theta_{\max} & \text{if } \langle v, Qv \rangle = 0. \end{cases} \end{aligned} \quad (11)$$

In this case linear convergence can be obtained from Theorem 4 and the error bound [2, Lemma 2.1]. However, the following more explicit rate of convergence can be obtained via a refinement of the proof of Theorem 4. In the particular case when  $Q$  is positive definite, Theorem 4 matches [14, Theorem 2].

**Theorem 5** *Assume the objective function  $f(y)$  in problem (8) is  $f(y) = \frac{1}{2} \langle y, Qy \rangle + \langle b, y \rangle$  where  $Q$  is an  $m \times m$  symmetric positive semidefinite matrix. Assume  $x_0 \in \Delta_{n-1}$  in Step 1 of Algorithm 1 is a vertex of  $\Delta_{n-1}$ . If the steplength  $\theta_k$  in Step 9 of Algorithm 1 is computed as in (11) then the convergence rate (9) holds with*

$$r = \frac{\Phi(Q^{1/2}A, Z^*)^2}{4\operatorname{diam}(Q^{1/2}A)^2} \geq \frac{\Phi(Q^{1/2}A)^2}{4\operatorname{diam}(Q^{1/2}A)^2}$$

provided  $Q^{1/2}A$  has at least two different columns.

**Proof:** This is a modification of the proof of Theorem 4. It suffices to show that when  $\theta_k = -\frac{\langle v, Qy_k + b \rangle}{\langle v, Qv \rangle}$

$$f(y_k) - f(y_{k+1}) \geq \frac{\Phi(Q^{1/2}A, Z^*)^2}{4\operatorname{diam}(Q^{1/2}A)^2} \cdot (f(y_k) - f^*). \quad (12)$$

Take  $z \in Z^*$  and  $y^* = Az$ . The inequality (12) in turn is a consequence of the following inequality:

$$\max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \langle Qy_k + b, a_\ell - a_j \rangle \geq \Phi(Q^{1/2}A, z) \cdot \sqrt{2(f(y_k) - f^*)}. \quad (13)$$

Indeed, equipped with (13) and proceeding as in the Premise 3 above, we get

$$\begin{aligned} -\langle Qy_k + b, v \rangle &\geq \frac{1}{2} \max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \langle Qy_k + b, a_\ell - a_j \rangle \\ &\geq \frac{\Phi(Q^{1/2}A, z)}{2} \cdot \sqrt{2(f(y_k) - f^*)}. \end{aligned}$$

Since  $\theta_k = -\frac{\langle v, Qy_k + b \rangle}{\langle v, Qv \rangle}$ , in particular  $\langle v, Qv \rangle > 0$ . The form of  $f$  and the choice of  $\theta_k$  then imply that

$$f(y_k) - f(y_{k+1}) = \frac{\langle v, Qy_k + b \rangle^2}{2 \langle v, Qv \rangle} \geq \frac{\Phi(Q^{1/2}A, z)^2}{4 \text{diam}(Q^{1/2}A)^2} \cdot (f(y_k) - f^*).$$

Since this holds for any  $z \in Z^*$ , (12) follows.

We now show (13). Let  $z \in Z^*$  and  $y^* = Az$ . We will assume  $Q(y_k - y^*) = QA(x_k - z) \neq 0$ . Below we deal with the case  $Q(y_k - y^*) = 0$ . Observe that

$$\langle Qy_k + b, y_k - y^* \rangle = f(y_k) - f^* + \frac{1}{2} \langle y_k - y^*, Q(y_k - y^*) \rangle. \quad (14)$$

Observe also that for any  $m \times m$  symmetric positive definite matrix  $M$  we have

$$\begin{aligned} & \max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \langle Qy_k + b, a_\ell - a_j \rangle \\ &= \max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \left\langle M^{-1/2}(Qy_k + b), M^{1/2}(a_\ell - a_j) \right\rangle \\ &\geq \Phi(M^{1/2}A, x_k, z) \left\langle M^{-1/2}(Qy_k + b), \frac{M^{1/2}(y_k - y^*)}{\|M^{1/2}(y_k - y^*)\|} \right\rangle \\ &= \Phi(M^{1/2}A, x_k, z) \cdot \frac{\langle Qy_k + b, y_k - y^* \rangle}{\|M^{1/2}(y_k - y^*)\|} \\ &= \Phi(M^{1/2}A, x_k, z) \cdot \frac{f(y_k) - f^* + \frac{1}{2} \langle y_k - y^*, Q(y_k - y^*) \rangle}{\|M^{1/2}(y_k - y^*)\|} \\ &\geq \Phi(M^{1/2}A, x_k, z) \cdot \frac{\sqrt{2(f(y_k) - f^*)} \|Q^{1/2}(y_k - y^*)\|}{\|M^{1/2}(y_k - y^*)\|}. \end{aligned}$$

The second step follows from the construction (1) of  $\Phi(M^{1/2}A, x_k, z)$  and the fact that by (14)

$$\left\langle M^{-1/2}(Qy_k + b), M^{1/2}(y_k - y^*) \right\rangle = \langle Qy_k + b, y_k - y^* \rangle > 0.$$

The fourth step follows from (14). The last step follows from the arithmetic-geometric inequality. Letting  $M \rightarrow Q$  and using that  $Q(y_k - y^*) = Q^{1/2}Q^{1/2}A(x_k - z) \neq 0$  we get

$$\begin{aligned} \max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \langle Qy_k + b, a_\ell - a_j \rangle &\geq \Phi(Q^{1/2}A, x_k, z) \cdot \sqrt{2(f(y_k) - f^*)} \\ &\geq \Phi(Q^{1/2}A, z) \cdot \sqrt{2(f(y_k) - f^*)}. \end{aligned}$$

Thus (13) is shown if  $Q(y_k - y^*) = QA(x_k - z) \neq 0$ . When  $Q(y_k - y^*) = QA(x_k - z) = 0$ , pick  $\tilde{z} \in \Delta_{n-1}$  sufficiently close to  $z$  such



that for  $\tilde{y} = A\tilde{z}$  both  $Q(y_k - \tilde{y}) \neq 0$  and  $\langle Qy_k + b, y_k - \tilde{y} \rangle > 0$  hold. This is possible because  $Q^{1/2}A$  has at least two different columns and  $\langle Qy_k + b, y_k - y^* \rangle > 0$  by (14). Proceeding as above with  $\tilde{z}$  and  $\tilde{y}$  in place of  $z$  and  $y$  we get

$$\max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \langle Qy_k + b, a_\ell - a_j \rangle \geq \Phi(Q^{1/2}A, \tilde{z}) \cdot \sqrt{2(f(y_k) - f(\tilde{y}))}.$$

Letting  $\tilde{z} \rightarrow z$  we get (13). ■

We should note some subtleties about  $\Phi(\cdot)$  relevant to Theorem 5. Observe that  $\Phi(A, x, z)$ ,  $\Phi(A, z)$ , and  $\Phi(A)$  depend on the set of atoms  $A$  and not on the potentially smaller set of vertices of  $\text{conv}(A)$ . A related subtlety is that although  $\Phi(A, x, z)$  and  $\Phi(A, z)$  are continuous on  $z$ , the quantities  $\Phi(A, x, z)$ ,  $\Phi(A, z)$ , and  $\Phi(A)$  are not continuous on  $A$ . Nonetheless, the condition  $Q(y_k - y^*) = Q^{1/2}Q^{1/2}A(x_k - z) \neq 0$  ensures that the limiting step  $M \rightarrow Q$  in the proof of Theorem 5 is sound. We also note that due to these subtleties, the value of  $\Phi(Q^{1/2}A)$  in Theorem 5 could be quite a bit smaller than  $\Phi(V)$  where  $V$  is the set of vertices of  $\text{conv}(Q^{1/2}A)$ .

On a related note, consider the special case  $f(y) = \frac{1}{2} \langle y, Qy \rangle$ . In this case the initial point  $x_0 \in \Delta_{n-1}$  can be chosen so that  $Q^{1/2}Ax_0$  is on the relative boundary  $\text{rbd}(\text{conv}(Q^{1/2}A))$  of  $\text{conv}(Q^{1/2}A)$ . For instance, pick  $\bar{x} \in \Delta_{n-1}$  such that  $QA\bar{x} \neq 0$  and start from  $x_0 = e_j$  for  $j = \text{argmin}_{i \in \{1, \dots, n\}} \langle QA\bar{x}, a_i \rangle$ . For any such initial point, all iterates  $y_k = Ax_k$  generated by Algorithm 1 have support contained in the set

$$B := Q^{1/2}A \cap \text{rbd}(\text{conv}(Q^{1/2}A)),$$

and consequently the bound in Theorem 5 can be improved to

$$r = \frac{\Phi(B, Z^*)^2}{4\text{diam}(B)^2} \geq \frac{\Phi(B)^2}{4\text{diam}(B)^2}.$$

## 5 Composite convex objective

To conclude, we extend the main ideas from Section 4 to the case when the objective is a composite function of the form

$$f(y) = g(Ey) + \langle b, y \rangle,$$

where  $E \in \mathbb{R}^{p \times m}$ ,  $b \in \mathbb{R}^m$ , and  $g : \mathbb{R}^p \rightarrow \mathbb{R}$  is a strongly convex function with Lipschitz gradient.

Consider problem (8) for this objective function. If  $L$  is an upper bound on the Lipschitz constant of  $\nabla g$ , then

$$f(y_k + \theta v) \leq f(y_k) + \theta \langle \nabla f(y_k), v \rangle + \frac{L\theta^2 \|Ev\|^2}{2}.$$

Consequently,  $\theta_k$  in Step 9 of Algorithm 1 could be computed as follows

$$\begin{aligned}\theta_k &:= \operatorname{argmin}_{\theta \in [0, \theta_{\max}]} \left\{ \langle \theta \nabla f(y_k), v \rangle + \frac{L\theta^2 \|Ev\|^2}{2} \right\} \\ &= \begin{cases} \min \left\{ \theta_{\max}, -\frac{\langle v, \nabla f(y_k) \rangle}{L\|Ev\|^2} \right\} & \text{if } Ev \neq 0 \\ \theta_{\max} & \text{if } Ev = 0. \end{cases} \quad (15)\end{aligned}$$

Once again, linear convergence can be obtained from Theorem 4 and the error bound [2, Lemma 2.1]. However, the more explicit rate of convergence in Theorem 6 holds.

**Theorem 6** *Assume in problem (8) the objective is  $f(y) = g(Ey) + \langle b, y \rangle$  where  $g$  is  $\mu$ -strongly convex and  $\nabla g$  is Lipschitz. Assume  $x_0 \in \Delta_{n-1}$  in Step 1 of Algorithm 1 is a vertex of  $\Delta_{n-1}$  and the steplength  $\theta_k$  in Step 9 of Algorithm 1 is computed as in (15) for some upper bound  $L$  on the Lipschitz constant of  $\nabla g$ . Then the convergence rate (9) holds with*

$$r = \frac{\mu}{L} \cdot \frac{\Phi(EA, Z^*)^2}{4\operatorname{diam}(EA)^2} \geq \frac{\mu}{L} \cdot \frac{\Phi(EA)^2}{4\operatorname{diam}(EA)^2}$$

provided  $EA$  has at least two different columns.

**Proof:** Assume  $p \leq m$ . If  $p > m$ , we can adapt the proof by taking a maximal subset of linearly independent rows of  $E$ .

The proof is a straightforward modification of the proof of Theorem 5. The crux is the following extension of (13):

$$\max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \langle \nabla f(y_k), a_\ell - a_j \rangle \geq \Phi(EA, z) \cdot \sqrt{2(f(y_k) - f^*)}. \quad (16)$$

To prove (16), first assume  $EA(x_k - z) \neq 0$ . Since  $g$  is  $\mu$ -strongly convex, we have

$$\langle \nabla f(y_k), y_k - y^* \rangle \geq f(y_k) - f^* + \frac{\mu}{2} \|E(y_k - y^*)\|^2. \quad (17)$$

Observe also that for  $\epsilon > 0$  sufficiently small the matrix  $M_\epsilon := \begin{bmatrix} E \\ 0 \end{bmatrix} + \epsilon I$  is non-singular. Using (17) and proceeding as in the proof of Theorem 5

we get

$$\begin{aligned}
& \max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \langle \nabla f(y_k), a_\ell - a_j \rangle \\
&= \max_{\ell \in I(x_k), j \in \{1, \dots, n\}} \langle M_\epsilon^{-1} \nabla f(y_k), M_\epsilon(a_\ell - a_j) \rangle \\
&\geq \Phi(M_\epsilon A, x_k, z) \left\langle M_\epsilon^{-1} \nabla f(y_k), \frac{M_\epsilon(y_k - y^*)}{\|M_\epsilon(y_k - y^*)\|} \right\rangle \\
&= \Phi(M_\epsilon A, x_k, z) \cdot \frac{\langle \nabla f(y_k), y_k - y^* \rangle}{\|M_\epsilon(y_k - y^*)\|} \\
&\geq \Phi(M_\epsilon A, x_k, z) \cdot \frac{f(y_k) - f^* + \frac{\mu}{2} \|E(y_k - y^*)\|^2}{\|M_\epsilon(y_k - y^*)\|} \\
&\geq \Phi(M_\epsilon A, x_k, z) \cdot \frac{\sqrt{2\mu(f(y_k) - f^*)} \|E(y_k - y^*)\|}{\|M_\epsilon(y_k - y^*)\|}.
\end{aligned}$$

Letting  $\epsilon \rightarrow 0$  and using that  $E(y_k - y^*) = EA(x_k - z) \neq 0$  we get (16). If  $EA(x_k - z) = 0$ , the above reasoning can be amended via a slight perturbation of  $z$  as in the proof of Theorem 5.  $\blacksquare$

## Acknowledgements

Javier Peña's research has been supported by NSF grant CMMI-1534850.

## References

- [1] S. Ahipasaoglu, P. Sun, and M. Todd. Linear convergence of a modified Frank-Wolfe algorithm for computing minimum-volume enclosing ellipsoids. *Optimization Methods and Software*, 23(1):5–19, 2008.
- [2] A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. Technical report, Faculty of Industrial Engineering and Management, Technion, 2015.
- [3] A. Beck and M. Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Math. Meth. of Oper. Res.*, 59(2):235–247, 2004.
- [4] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.
- [5] M. Epelman and R. M. Freund. Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system. *Math. Program.*, 88(3):451–485, 2000.
- [6] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Quarterly*, 3:95–110, 1956.

- [7] D. Garber and E. Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. Technical report, Faculty of Industrial Engineering and Management, Technion, 2013.
- [8] J. Guélat and P. Marcotte. Some comments on Wolfe’s away step. *Math. Program.*, 35:110–119, 1986.
- [9] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, volume 28 of *JMLR Proceedings*, pages 427–435, 2013.
- [10] P. Kumar and E. A. Yildirim. A linearly convergent linear-time first-order algorithm for support vector classification with a core set result. *INFORMS Journal on Computing*, 23(3):377–391, 2011.
- [11] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- [12] R. Nanculef, E. Frandi, C. Sartori, and H. Allende. A novel Frank-Wolfe algorithm. Analysis and applications to large-scale SVM training. *Inf. Sci.*, 285:66–99, 2014.
- [13] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Kluwer Academic Publishers, 2004.
- [14] J. Peña, D. Rodríguez, and N. Soheili. On the von Neumann and Frank-Wolfe algorithms with away steps. *SIAM J. on Optim.*, To Appear, 2015.
- [15] P. Wolfe. Convergence theory in nonlinear programming. In *Integer and Nonlinear Programming*. North-Holland, Amsterdam, 1970.